Structural principles of prokaryotic gene regulatory proteins and the evolution of repressors and gene activators

Gali Prag [1], Simon Greenberg [2] and Amos B. Oppenheim[*]

[1]Department of Molecular Genetics and Biotechnology and the [2]Computers unit, The Hebrew University - Hadassah Medical School P. O. Box 1172, Jerusalem, 91010, Israel

[*]Tel: 972-2-6757309, Fax: 972-2-6757308

E-mail: ao@cc.huji.ac.il

Mol Microbiol 1997 Nov;**26** 619

Gene regulators are guided, in bacteria, to specific DNA sequences by a helix-turn-helix (HTH) motif found tethered to larger, regulatory, response domains. The DNA binding domain is most often a compact HTH motif of about 20 amino acid residues. This domain selects DNA binding sites in close proximity to a promoter with high affinity, as has been clearly shown for phage repressors and the Lac-Gal repressor family (Weickert and Adhya, J. Biol. Chem. 267:15869-74, 1992). We analyzed all the available (74) annotated E. coli repressors and activators protein sequences present in the SwissProt bank (Bairoch and Apweiler, 1996, Nucleic Acids Res. 24: 21-25) for the position of the HTH motif and uncovered a number of general structural rules for bacterial regulatory proteins. Dedicated repressors transcriptional activators were analyzed separately. Proteins with dual function were included in the latter group. The analysis of 30 E. coli repressor proteins, encompassing a number of protein families, showed that with the exception of two cases, TrpR and

CelD, all carry the HTH motif at the amino terminus (Fig. 1A). TrpR, a small 107 residue long protein we suspect is evolutionarily distinct from the other repressor proteins. CelD belongs to the AraC protein family, of which all other members are transcriptional activators. We predict that CelD may also function as a transcriptional activator and should perhaps be included as a dual function regulator in the second group.

The analysis of 44 E. coli activators and dual regulators (Gralla and Collado-Vides, 1996, In *Escherichia coli* and *Salmonella* cellular and molecular biology, F. C. Neidhardt et al (eds) ASM Press, pp.1232-1245) revealed that 21 carry the HTH motif at the N-terminus (Fig. 1B). The remaining can be divided to two classes. One class of 16 members, of various sizes, were found to carry the HTH motif at the extreme C-terminus. For example, the HTH motif in the 210 amino acid long CRP is found at residues 170-189, in the 216 amino acid long NARL at residues 173-192, in the 441 amino acid long HYDG at residues 421-440, in the 469 amino acid long NTRC at residues 445-464, in the 513 amino acid long TYRR at residues 483-502, in the 692 amino acid long FHLA at residues 663-682 and at residues 853-872 in the 901 amino acid long MALT. The remaining 7 members are proteins belonging to the AraC family that carry two putative HTH domains at their C-terminus. In some of these it is not known which HTH motif is the DNA binding domain. In MELR the more distant HTH, which is highly conserved in this class, plays a major role in gene activation (Caswell *et al*., 1992, Biochem. J. 287:493 9.), while in ARAC the internal HTH motif is more critical (Schleif, 1996, In *Escherichia coli* and *Salmonella* cellular and molecular biology. F. C. Neidhardt et al (eds). ASM Press, pp.1300-1309).

The following general structural rules for prokaryotic regulatory proteins carrying a HTH motif emerge from our limited study with a sample of *E. coli* proteins:

i. In repressor molecules the HTH motif is present at the N-terminus.

ii. In activator proteins and in dual function regulators the HTH motif is present at either the N- or the C-terminus.

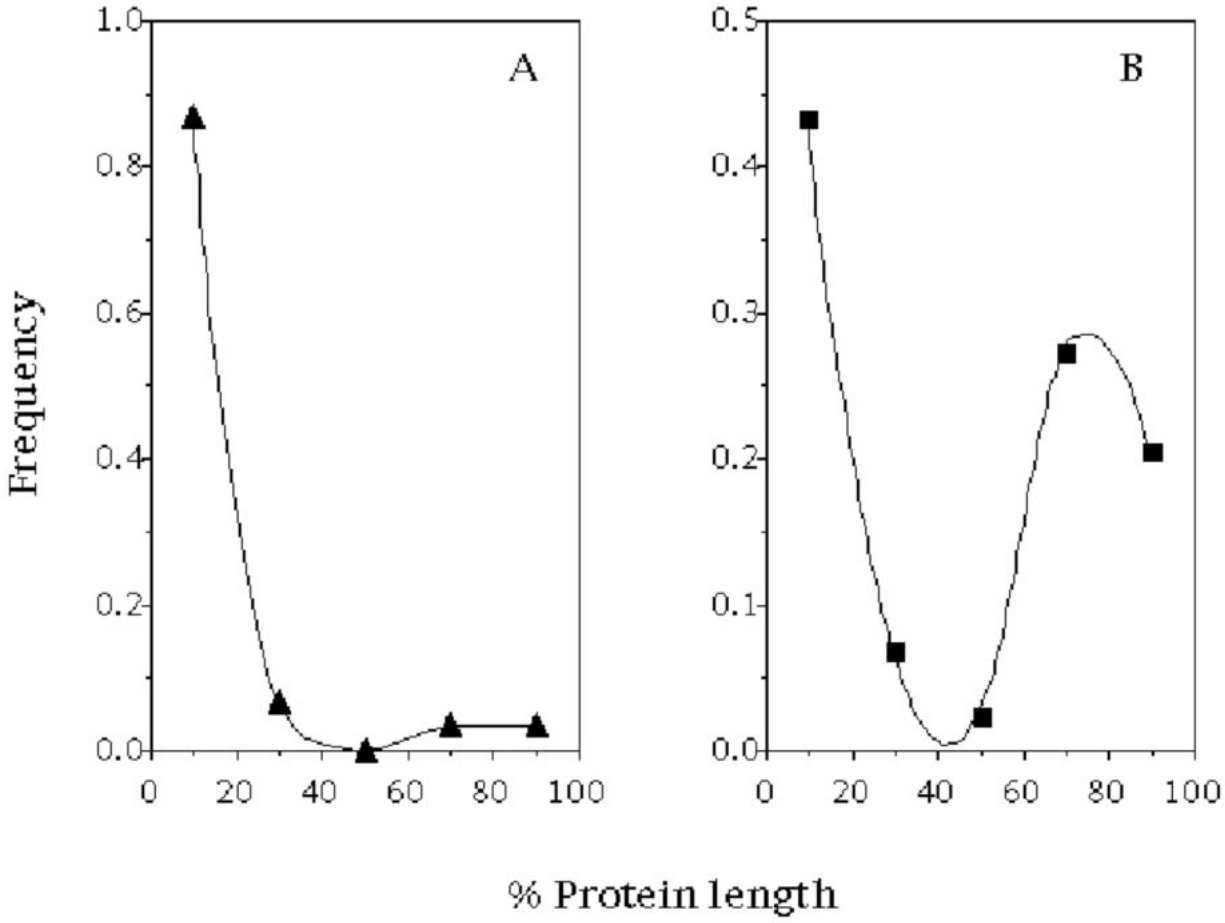

Fig. 1. The position of HTH motif in E.coli regulatory proteins.

The SwissProt database (Oct. 3ed, 1996) was extracted for all prokaryotic DNA binding domain proteins carrying an annotated HTH motif using standard procedures of the GCG Wisconsin program package. These annotations are based on experiments evidence and amino acid similarities. For each protein the position of the amino terminal residue of the HTH motif divided by the total number of amino acid residues was calculated as the location of the HTH motif along the protein, as % of protein length. These numbers allowed us to group the proteins into 5 groups (1-20%, 21-40% etc.) and the results are presented as frequency in a collection of 30 repressor proteins (A) and 44 activators and dual function proteins (B). Repressors: ACRR, AGAR, ARSR, ASCG, BETI, BIRA, CELD, CSCR, CYTR, DEOR, DICR, DICC, EBGR, ENVR, FRUR, GALR, GALS, GATR, GLPR, HIPB, ICLR, LACI, LEXA, MALI, PDHR, PURR, RBSR, SRLR, TRER, TRPR. Activators with HTH motif at the N-terminus: ALPA, ASNC, CYNR, CYSB, DSDC, FADR, FUCR, GCVA, GUTM, ILVY, LEUO, LYSR, MARA, METR, NAGC, NHAR, OXYR, SOXR, SOXS, TDCA, XAPR. Activators with HTH motif at the C-terminus: ATOC, CRP, EVGA, FHLA, FNR, HYDG, MALT, NARL, NARP, NLI NTRC, RCSA, RCSB, SDIA, TYRR, UHPA. Proteins with HTH motif at the C-terminus belonging to the AraC family: ADA, APPY, ARAC, ENVY, MELR, RHAR, RHAS.

We propose that this unique organization results from the different nature and function of the primitive regulators and reflects evolutionary events rather than from the random addition of a HTH motif at the N- or C-terminus. We postulate that the primitive repressor was made of a DNA binding HTH motif, sufficient to block RNA polymerase action. Fusions between DNA sequences coding for

a HTH motif and a response motif gave rise to ancestral repressors, endowed with the ability to respond to environmental changes and thus confer better fitness upon the cell. The response domains evolved independently from ancestral proteins. For example, the inducer binding domain of the Lac repressor is a homolog of the bacterial periplasmic binding proteins participating in sugar transport (Mowbray and Cole, 1992, J. Mol. Biol. 225:155-75).

Genes coding for regulators are generally located in monocistronic transcription units, driven by weak promoters, presumably in order to maintain a low level of these regulators. We propose that response domains were added to the region coding for the HTH repressor domain at its C-terminus, without disrupting the association of the primitive repressor with its promoter and control elements. This only requires that the two open reading frames are fused in phase, avoiding the addition of a sequence into a region crowded with transcription and translation signals. Further evolution probably occurred by gene duplication and mutational events that also led to the emergence of activators which carry the HTH motif at the N-terminus. Activators that bind upstream of a promoter can, for example, by evolving small protein surfaces, recruit RNA polymerase (increasing KB) by specific protein-protein contacts. In contrast, we suggest that activators bearing a HTH motif at the C-terminus evolved from primitive activators which did not possess a DNA binding motif. This class of activators act mainly by stimulating the isomerization of RNA polymerase from the closed to the open complex (increasing kf) by a yet unknown mechanism. It was shown that activators acting on s54-RNA polymerase can function, although at a lower efficiency, following deletion of the HTH motif. (Berger, *et al,* 1995 J. Bacteriol. 177:191-9). The HTH motif allows for specific binding to enhancers and thereby improve their specificity. We suggest that the preferred way of acquiring a HTH motif was by recombination at the C-terminus of the primitive activator. It should be pointed

out that CRP, the best studied transcriptional activator, can activate different promoters by increasing either KB or kf (Busby and Kolb, 1996, In Regulation of gene expression in *Escherichia coli*, E. C. C. Lin et al (eds) Chapman and Hall, pp 255-279) suggesting that CRP was subjected to additional evolutionary events.

Fusions between domains in prokaryotic proteins was previously suggested as for example, for proteins carrying fibronectin type III (Fn3) domains (Bork and Doolittle, 1992, Proc. Natl. Acad. Sci. USA 89: 8990-4). The presence of well characterized groups of regulators such as the Lac-Gal and LysR family with an N-terminal HTH and the NTRC and Crp family with a C-terminal HTH supports the notion of multiple fusion events between DNA sequences coding for the HTH motif and the response motifs. It should be noted that, in general, in the large number of enzymes manipulating DNA, one does not find a distinct separation between the DNA binding domain and the active site domain (for references see (Riley and Labedan, 1996, In Escherichia coli and Salmonella cellular and molecular biology, F. C. Neidhardt et al (eds) ASM Press, pp. 2118-2202)). This is true not only for the non-specific DNA binding enzymes such as DNA polymerases and topoisomerases but also for those enzymes, that bind to specific sites such as restriction endonucleases and DNA methylases. We suggest that the genes coding for these enzymes were evolved independently from gene regulators, originating from different ancestral genes. In addition, nucleoid- proteins, such as HU, IHF and H-NS that can regulate gene expression, probably evolved independently of the regulatory proteins analyzed above.